Amendments to the Specification

Please replace the section beginning at page 3, line 12, with the following rewritten section:

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the present invention and for further features and advantages, reference is now made to the following description, taken in conjunction with the accompanying drawings, in which:

FIGURE 1 is a flowchart illustrating an example method for generating a microarray;

FIGURE 2 illustrates background noises and includes FIGURES 2(a), 2(b), and 2(c);

FIGURE 3 illustrates noise settings and includes FIGURES 3(a) and 3(b);

FIGURE 4 illustrates an example cDNA microarray spot model;

FIGURE 5 illustrates variability in spot size and spread from the spot size and includes FIGURES 5(a), 5(b), and 5(c);

FIGURE 6 illustrates inter-spot grid spacing and includes FIGURES 6(a), 6(b), and 6(c);

FIGURE 7 illustrates an effect of a radius drift variation and includes FIGURES 7(a), 7(b), and 7(c);

FIGURE 8 illustrates chord rate settings and includes FIGURES 8(a), 8(b), and 8(c);

FIGURE 9 illustrates an edge noise of spots and includes FIGURES 9(a), 9(b), and 9(c);

FIGURE 10 illustrates fluorescent detection response characteristic functions and includes FIGURES 10(a), 10(b), 10(c), and 10(d);

FIGURE 11 illustrates possible scatter plots due to various response conversions for different fluorescent channels and includes FIGURES 11(a), 11(b), and 11(c);

FIGURE 12 illustrates increased spike noise levels and includes FIGURES 12(a), 12(b), and 12(c);

FIGURE 13 illustrates scratch noise and parameter settings and includes FIGURES 13(a), 13(b), and 13(c);

FIGURE 14 illustrates parameter settings for snake noise and includes FIGURES 14(a), 14(b), and 14(c);

FIGURE 15 illustrates convolution kernels and includes FIGURES 15(a) and 15(b);

FIGURE 16 illustrates a three-dimensional profile before and after smoothing and includes FIGURES 16(a), 16(b), and 16(c);

FIGURE 17 illustrates a full size array simulation with different parameter settings and includes FIGURES 17(a) and 17(b);

FIGURE 18 illustrates a comparison between a simulated signal versus an extracted signal from a microarray image analysis program and includes FIGURES 18(a) and 18(b); and

FIGURE 19 illustrates simulated images exhibiting undesirable noise conditions and includes FIGURES 19(a), 19(b), 19(c), and 19(d).

Please replace the paragraph beginning at page 1, line 14, with the following rewritten paragraph:

Since the inception of cDNA microarray technology [1] as a high throughput method to gain information about gene functions and characteristics of biological samples, many applications of the technology have been reported [2-10]. With the improvement of the technology, including fabrication, fluorescent labeling, hybridization, and detection, many computer software packages for extracting signals arising from tagged mRNA hybridized to arrayed cDNA locations have been designed and applied in various experiments ~~[11-13]~~ [11-12]. As reported in [11], a target detection procedure has been implemented that utilizes manually specified target arrays, extracts the background

via the image histogram, predicts target shape by mathematical morphology, and then evaluates the intensities from each cDNA location and its corresponding ratio quantity.

Please replace the paragraph beginning at page 7, line 27, with the following rewritten paragraph:

Rather than be constant across the entire microarray, the mean of the background noise may vary owing to various scanning effects. It can take different shapes: parabolic, positive slope, or negative slope. In this case a function $g(x, y)$ is first generated (parabolic, positive slope, or negative slope) to form a background surface and normal noise is added to it pixel wise. Thus, the background intensity is of the form $I_b \sim N(\mu_b, \sigma_b^2)$ with $\mu_b = \gamma g(x, y)$, where $\gamma \sim U(a, b)$ is the targeted background noise level. Background deviation is set independently for each channel: $\sigma_{b_1} = k_{b_1}\mu_b$ and $\sigma_{b_2} = k_{b_2}\mu_b$. FIGURE 2 shows various ~~noise backgrounds~~ background noises with the background deviation factor set at $k_{b_1} = k_{b_2} = 0.1$ *(10%)*. The signal-to-noise ratio (SNR) is set at 1.0 for the slides of FIGURE 2. The slides have the following settings: FIGURE 2(a), Parabolic background noise, FIGURE 2(b), Positive slope background, FIGURE 2(c), Negative slope background all with global noise parameter.

Please replace the paragraph beginning at page 8, line 4, with the following rewritten paragraph:

In many practical examples, the non-specific hybridization at the target location may be different from its peripheral region. Although one may have trouble pin-pointing this particular observation under normal conditions owing to signal interference, it is sometimes unmistakable when locations assumed to be weakly expressed, or not expressed at all, carry some non-zero readouts, or the intensity in the center is stronger than the doughnut-ring if the printed target is doughnut-shaped. This artifact is simulated under a gradient noise condition by allowing the background for the center holes to be at

higher levels than the signal intensities. Hence, there is an option to use global background or local background information to set the noise parameter for the center hole. FIGURE 3 shows the effects of using local and global background parameters. The examples of FIGURE 3 show different noise settings for a spot's inner hole. FIGURE 3(a) uses a global background parameter to fill the center hole. FIGURE 3(b) uses local background for filling the center hole. The background noise is set to sloped type with SNR of 1.5. This effect may not appear everywhere in a simulated image; however, it is often sufficient to require appropriate procedure design in the image analysis program to lessen the penalty. The effects of weak targets will be further studied in later sections.

Please replace the paragraph beginning at page 9, line 8, with the following rewritten paragraph:

Prior to distortion and noise, the cDNA deposition spot is considered to be circular with random radius S. The mean of the radius is set according to the array density and its variance relates to the consistency of spot size. S is modeled by a normal distribution having mean μ_s and variance σ_s^2, $S \sim N(\mu_s, \sigma_s)$, with the standard deviation being a pre-determined proportion, k_s, of the mean, or $S \sim N(\mu_s, k_s\mu_s)$. The radius mean is set for every block, and randomized over a small range within the array. The block randomness of μ_s is modeled by a uniform distribution, $\mu_s \sim U(s_a, s_b)$. FIGURE 5 shows parts of blocks with spot radii depending on the number of spots in a block. ~~For FIGURES 5a through 5c, the block portions are for block sizes (10, 15), (25, 45), and (25, 45), respectively~~ FIGURES 5(a), 5(b), and 5(c) show the variability in spot size and spread from its size. The spot radius distribution is automatically set depending on the number of spots in a block (width, height), where (*col, row*) denotes the number of columns and rows within the block, respectively. In FIGURE 5(a), the block portion is for size (10,15), μ_s~U[23.3 24.3]. In FIGURE 5(b), the block portion is for size (20, 25) , μ_s ~U[12.6 13.6]. In FIGURE 5(c), the block portion is for size (25, 45) , μ_s~U[5.45 6.45]. Standard deviation k_s, equals 1%, 7%, and 20% of radius, respectively.

Occasionally, a spot overlaps with it neighbors ~~(FIGURE 5c)~~ (FIGURE 5(c)) when k_r, k_s is set to a larger proportion. This situation simulates the condition where too much cDNA solution is deposited and/or the drying process may be slow in comparison to the liquid spreading process.

Please replace the paragraph beginning at page 9, line 22, with the following rewritten paragraph:

Depending on the robot arm and printing ability of the pins, the inter-spot distance, G_{sp}, may vary. Owing to the physical mechanics of the robot arm, the block size (pixel units) is fixed in most cases. The inter-spot distance can be set to accommodate spot size and random variation in spot radii. The effects are illustrated in FIGURE 6, where the number of rows and columns are fixed. FIGURE 6 shows inter-spot grid spacing. FIGURE 6(a) shows G_{sp} = 3 pixels, μ_s~U[9.5 10.5]. FIGURE 6(b) shows G_{sp} = 6 pixels, μ_s~U[8 9]. FIGURE 6(c) shows G_{sp} = 10 pixels, μ_s~U[6.5 7.5]. The example shown has (35, 20) rows, columns respectively with k_s = 0.05.

Please replace the paragraph beginning at page 10, line 8, with the following rewritten paragraph:

Some microarray scanners capture two fluorescent signals in two passes of scanning. Due to the mechanical homing error, the two fluorescent channels may not align exactly. In these settings, some small offset between the two channels can be observed. This offset may occur at sub-pixel resolution. To simulate this offset, the model offers a random offset between the centers of the two channels. It is achieved by randomly offsetting the spot center of the second channel by one pixel in either of the horizontal and vertical directions. These offsets are applied following application of the spot drifts in the first channel. FIGURE 7 illustrates the spot drift. FIGURE 7 shows the effect of radius drift variation (P_D, R_{d1}, R_{d2}). The settings for FIGURE 7(a) are (0.05, 5, 100). The settings for FIGURE 7(b) are (0.25, 15, 100). The settings for FIGURE 7(c)

are (0.5, 50, 100). As the activation probability with drift range is set higher, a spot drifts away from its center.

Please replace the paragraph beginning at page 12, line 1, with the following rewritten paragraph:

Once the number of chords for a spot is determined, the distance, L, of each chord center to the edge is selected from a beta distribution: $L \sim B(\alpha_L, \beta_L)$. Inter-block variability is modeled by allowing α_L and β_L to be randomly selected from uniform distributions: $\alpha_L \sim U(a_\alpha, b_\alpha)$, and $\beta_L \sim U(a_\beta, b_\beta)$. Owing to the large family of shapes generated by beta distributions, this provides a wide range of distributions for L. Finally, the chord locations are chosen uniformly randomly according to an angle $\theta \sim U(0, 2\pi)$. FIGURE 8 illustrates the effect of selecting increased chord rates. FIGURE 8 shows different chord rate settings for each of the slides. The probability weights for (0,1,2,3,4) chords were set at the following levels: FIGURE 8(a) $p_0 = 0.70$, $p_1 = 0.30$; $p_2 = 0$; $p_3 = 0$; $p_4 = 0$; FIGURE 8(b) $p_0 = 0.20$, $p_1 = 0.40$, $p_2 = 0.25$, $p_3 = 0.15$; $p_4 = 0$; FIGURE 8(c) $p_0 = 0$, $p_1 = 0.10$, $p_2 = 0.40$, $p_3 = 0.30$, $p_4 = 0.20$. The chord rate is reset at the beginning of a block.

Please replace the paragraph beginning at page 13, line 9, with the following rewritten paragraph:

where δ controls the threshold and hence the edge noise, and Δ denotes the symmetric difference. δ is used as controlling parameter. S' is a binary mask giving the spatial domain of the spot. FIGURE 9 shows edge noise for various δ thresholds. FIGURE 9 shows the edge noise on the spots. A noise controlling parameter (δ) can be set from [0, 1.0]. The examples of FIGURE 9 show an increased edge noise effect. For FIGURE 9(a), $\delta = 0.25$. For FIGURE 9(b), $\delta = 0.1$. For FIGURE 9(c), $\delta = 0.03$.

Please replace the paragraph beginning at page 13, line 20, with the following rewritten paragraph:

It is well known that the distribution of gene expression levels within a cell closely follows an exponential distribution ~~[26]~~ [13]. Given a microarray containing N genes, the intensity levels I_k, for $k = 1, \ldots, N$, assumed to be related to the expression levels of N genes, are simulated by an exponential distribution. This intensity level I_k is considered to be the ground-truth signal that is not directly measurable from microarray, since from either biological or bio-chemical processes, from mRNA extraction up to the hybridization process, some variation will be introduced into measurement of final fluorescent signal strength. For each microarray, a particular exponential distribution with mean β is first chosen (for a detection system with gray-level up to 65,535, β is usually selected around 3000). Then at each spot location, which is assumed to represent one unique gene, one ground-truth signal level I_k is generated from the exponential distribution. For two observable measurements (R_k, G_k) from two fluorescent channels, two numbers are generated from a normal distribution with mean of I_k and standard deviation of αI_k, where α is a pre-determined coefficient of variation, which is usually about 5% to 30% depending on the assumed biological relation between the two channels.

Please replace the paragraph beginning at page 15, line 3, with the following rewritten paragraph:

Owing to various reasons, such as imprecise quantities of starting mRNA for the two channels, different labeling efficiencies, or uneven laser powers at the scanning stage, in actual microarray experiments there may not be equal intensities even if two channels use exactly the same labeled mRNA. Moreover, one may not be able to assume that the fluorescent intensity is linearly related to the expression level. In fact, it is very difficult to determine the exact form of the response function from expression level to intensity due to the complex combination of bio-chemistry to photon-electronics. A family of functions that covers most of the understandable conditions, shown in FIGURE

10, such as delayed response, saturation (which is an embedded feature in the digital system since in general no gray-level can pass 16-bit binary digits in a typical microarray system), and unbalanced channel intensity, is selected. FIGURE 10 shows fluorescent detection response characteristic functions. In the figures, the middle (circled) curve is the reference function with the parameters of $(a_0, a_1, a_2, a_3) = (0, 100, -1, 1)$. Also, in all the figures, the x-axis is the input signal intensity, and y-axis is the observed signal intensity, and both are in $\log_{10}$-scale. FIGURE 10(a) shows delayed response at various levels, with fixed $a_0 = 0$ and $a_3 = 1$. FIGURE 10(b) shows different amplification levels, with fixed $a_0 = 0$ and $a_2 = -1$. FIGURE 10(c) shows different response curvature, with fixed $a_0 = 0$ and $a_3 = 1$. FIGURE 10(d) shows some other parameter settings, with fixed $a_3 = 1$. This simulation is intended to facilitate understanding as to what is the best way for expression ratio normalization, whether linear based methods will be sufficient or non-linear based methods may be needed. The function family is characterized by four parameters, (a_0, a_1, a_2, a_3), and the function form is given by

Please replace the paragraph beginning at page 16, line 15, with the following rewritten paragraph:

The scatter plots in FIGURE 11 show the effects of the channel normalization. By choosing different parameter sets, one can simulate many of the situations observed in real microarray images. FIGURE 11 shows possible scatter plot due to various response conversions for different fluorescent channels. 10,000 data points (gene expression levels) were generated by the exponential distribution with mean of 3000. After passing through two fluorescent channels (with some response characteristic functions as shown in FIGURES 11(a) to 11(c)), data variations were added by passing each data point through a normal distribution with the standard deviation to be 15% of mean expression signal. FIGURE 11(a) shows without any alteration (or equivalently, set parameters for the response function to be $(a_0, a_1, a_2, a_3) = (0, 1, -1, 1)$), and assume the signal intensities from red channel and green channel are equivalent (a simulated self-self experiment). FIGURE 11(b) shows a banana shape. Intensity in the green channel pass a response

function with parameters (a_0, a_1, a_2, a_3)= (0, 500, -1, 1), where the red channel takes the parameters (0, 10, -1, 1). FIGURE 11(c) shows a sinusoid shape. The red channel's response function with parameters (0, $100^{1/0.7}$, -0.7, 1) and the green channel with (0, $100^{1/0.9}$, -0.9, 1).

Please replace the paragraph beginning at page 17, line 7, with the following rewritten paragraph:

In a practical biology laboratory, it is not necessary to maintain a dust-free environment. Hence, fine microscopic dust particles are nearly impossible to avoid. On laser excitation, these particles fluoresce to give high intensity spikes. Moreover, in some cases, bad mixtures of cDNA solutions result in precipitation, and these particles fluoresce with a very high intensity. These effects are simulated by adding spike noise at a preset rate. Such intensity spikes are added randomly across the entire slide area, the number of such noise pixels being preset in terms of the total number of pixels in the array. The amount of spike noise in an array is set with reference to the percentage, L_{spi}, of the total number of pixels in the array. Typical low to high noise levels are be set by selecting 0.1% to 10%. Once a pixel is selected for spike noise, the adjacent pixels have a higher probability of being affected. Thus, a random number, W_{spi}, of pixels are chosen in an arbitrary direction to be influenced by this noise. The intensity, N_S, of the spike noise is governed by an exponential distribution with mean μ_{spi}. In FIGURE 12, the exponential mean is fixed but the spike rate is increased through the parts of the figure. FIGURE 12 shows increased spike noise levels L_{spi}. FIGURE 12(a) shows a level of 0.1%. FIGURE 12(b) shows a level of 5%. FIGURE 12(c) shows a level of 10%. An exponential rate range is maintained for the above.

Please replace the paragraph beginning at page 17, line 24, with the following rewritten paragraph:

Physical handling of the array slides can result in surface scratches. These typically result in low intensity levels. Scratch-noise intensity is parameterized as a ratio,

κ_{sc}, giving the background-to-scratch-noise intensity level. Other parameters are the number of strips, strip thickness W_{sc}, and a random strip length, L_{sc}, given as a multiple of the spot size. The latter is modeled as a uniform distribution: $L_{sc} \sim U[L_{sc1}, L_{sc2}]$. Strips are placed at random positions on the array, and are inclined according to a (discrete) uniformly random angle, $\theta_{sc} \in \{0^o, 45^o, 90^o, 135^o, 180^o\}$. FIGURE 13 shows scratch noise with its parameter settings; the number of strips is fixed at 7. FIGURE 13 shows the noise for incremental parameter settings: ~~(a) $L_{sc} \sim U[2, 7]$, $\kappa_{sc} = 2.0$, $W_{sc} = 4$ pixels; (b) $L_{sc} \sim U[5, 10]$, $\kappa_{sc} = 3.0$, $W_{sc} = 7$ pixels; (c) $L_{sc} \sim U[7, 15]$, $\kappa_{sc} = 4.0$, $W_{sc} = 10$ pixels. The number of strips is fixed at 7.~~ FIGURE 13(a) shows $L_{sc} \sim U[2\ \ 7]$, $\kappa_{sc} = 1.5$, $W_{sc} = 3$ pixels; FIGURE 13(b) shows $L_{sc} \sim U[5\ \ 15]$, $\kappa_{sc} = 2.5$, $W_{sc} = 7$ pixels; FIGURE 13(c) shows $L_{sc} \sim U[8\ \ 45]$, $\kappa_{sc} = 4.0$, $W_{sc} = 15$ pixels. The number of scratches is maintained to 7. The noise factor $\kappa_{sc} = 0.1$.

Please replace the paragraph beginning at page 18, line 5, with the following rewritten paragraph:

Fine fabric dust particles on the slides can create snake-tailed strips on laser excitation. These strips are normally higher intensity than the signal level. To simulate this noise, an equiprobable multi-directional snake noise has been generated consisting of some number, N_{seg}, of segments. Analogously to scratch noise, the intensity parameterized as a ratio, κ_{sn}, giving the average-signal-to-snake-noise intensity level, the number of snakes, snake thickness W_{sn}, and a random length, L_{sn}, given as a multiple of the spot size. The latter is modeled as a uniform distribution: $L_{sn} \sim U[L_{sn1}, L_{sn2}]$. FIGURE 14 shows different parameter settings for snake noise. FIGURE 14 shows the noise for incremental parameter settings: FIGURE 15(a) shows $N_{seg} = 5$, $L_{sn} \sim U[5, 10]$, $\kappa_{sn} = 0.50$, $W_{sn} = 2$ pixels; FIGURE 15(b) shows $N_{seg} = 10$, $L_{sn} \sim U[5, 30]$, $\kappa_{sn} = 0.33$, $W_{sn} = 3$ pixels; FIGURE 15(c) shows $N_{seg} = 15$, $L_{sn} \sim U[~~15~~ 5, 80]$, $\kappa_{sn} = 0.25$, $W_{sn} = 5$ pixels. The direction of the tail was randomly chosen with equal probability for each.

Please replace the paragraph beginning at page 24, line 29, with the following rewritten paragraph:

FIGURE 17 shows two microarrays generated with $NS_w = 35$ rows and $NS_h = 25$ columns, at $B_h = B_w = 900$ pixels per block. The example shows full size arrays simulation with different parameter settings. Depending on the parameters the arrays are called as "average" and "noisy" in quality. FIGURE 17(a) shows good quality, has SNR of 2.0, with normal background, and spike noise $L_{spi} = 0.3\%$. Array boundaries are set at $(M_t, M_l, M_r, M_b) = (100, 100, 100, 100)$. By choosing parameters, two different array qualities have been generated. ~~Part a~~ FIGURE 17(a) illustrates an ideal microarray image with normal background and parameters $\beta = 3000$, $SNR = 2.0$, $\alpha = 0.05$, $G_{sp} = 6$, $P_D = 0.05$, $(d_a, d_b) = (2, 15)$, $(k_{b1}, k_{b2}) = (10, 10)$, $P_{outlier} = 0.05$, $L_{spi} = 0.3\%$, $\delta_{ed} = 0.3$,

Please replace the paragraph beginning at page 25, line 13, with the following rewritten paragraph:

FIGURE 17(b) shows a noisy array with SNR of 1.1 with parabolic background noise, and spike noise $L_{spi} = 15\%$. ~~Part~~ FIGURE 17(b) illustrates a noisy microarray image with parabolic background and parameters: $\beta = 3000$, SNR = 1.1, $\alpha = 0.25$, $G_{sp} = 4$, $P_D = 0.4$, $(d_a, d_b) = (15, 100)$, $(k_{b1}, k_{b2}) = (25,2\ 5)$, $P_{outlier} = 0.7$, $L_{spi} = 15\%$, $\delta_{ed} = 0.03$,

Please replace the paragraph beginning at page 25, line 26, with the following rewritten paragraph:

To illustrate how the simulation can be used to analyze microarray image software, the ArraySuite [11] software is applied to extract the image intensities and ratios from the image and then these are compared to the corresponding intensities and ratios used for simulation. The ideal case is used to illustrate the utility of the simulation. FIGURE 18 shows a comparison between simulated signal (ideal setting) vs. extracted signal from a microarray image analysis program. In FIGURE ~~18a~~ 18(a), intensities from one fluorescent channel have been extracted (y-axis) and plotted against the simulation

signal intensities. FIGURE 18(a), shows signal extracted from one fluorescent channel (y-axis) comparing to the signal used for simulation in the same channel (x-axis). The extracted signal generally corresponds well to the simulated signal, with some variation. After excluding intensities less than 300, the mean and standard deviation of the difference between the two $\log_{10}$-transformed intensities are 0.016 (or $10^{0.016} = 1.038$) and 0.038 (or $10^{0.038} = 1.09$), respectively. The ratio comparison is given in FIGURE ~~18b~~18(b). FIGURE 18(b) shows ratio from microarray image analysis program (y-axis) comparing to the ratios generated by the simulation (x-axis). When signal intensity is weak (less than 300), various noise components in the simulation process affect the accuracy of the signal extraction program. Since the problem may be unavoidable, a measurement quality metric may be used to provide confidence in downstream data analysis. In this case, if the signal intensity is less than 300, then the noise interaction is significant.

Please replace the paragraph beginning at page 26, line 19, with the following rewritten paragraph:

The simulation program has been used extensively in the design of the microarray image-analysis program used at the National Human Genome Research Institute. This has been done by testing the accuracy of the analysis program on simulated images exhibiting troublesome noise conditions and then tuning the program to achieve better results. One such application concerns large and overlapping spots, as illustrated in FIGURE 19(a), which shows part of an actual hybridized image in which some spots are substantially larger than intended owing to randomness in the cDNA deposition procedure. FIGURE 19(a) shows part of an actual hybridized image with spots larger than average. This defect causes various problems, one being poor background estimation. This problem is illustrated by simulating an image with large spot size variation and drifting conditions [FIGURE 19(b)]. FIGURE 19(b) shows a simulated microarray with larger spots and spots overlapping with their neighbors. If the image analysis program extracts the local background by averaging the region around the bounding box (which was used as a

starting condition in an earlier version of the NHGRI program), an elevated background average may be obtained since the bounding box may overlap neighboring targets that are large in size and strong in expression level. An additional problem is that some weak targets may not be detected [FIGURE 19(c)]. FIGURE 19(c) shows that an original background intensity extraction program produces an undetected spot (target in the middle without any outer boundary). Based on these considerations, the program has been modified to calculate the four average intensities from the four corners and the four average intensities from the four sides of the bounding box, and then take the minimum among these as the initial estimation of the local background. A histogram-based method is then invoked around the initial estimated background to further improve the estimation [11]. The output from FIGURE 19(b) according to the modified program is shown in FIGURE 19(d): the weak target is detected and there is improved local background estimation for the spots. FIGURE 19(d) shows that the improved background extraction program more accurately measures the local background intensity and effectively allows detection of weak targets.

Please replace the paragraph beginning at page 28, line 29, with the following rewritten paragraph:

13. ~~See www.imgresearch.com, genome-www.stanford.edu/microarray, www.axon.com, www.imagingresearch.com and www.nutecsciences.com.~~ Bishop, J. O., J.G. Morton, et al., Three abundance classes in Hela cell messenger RNA. 1974, *Nature*, **250**(463): 199-240

Please delete the paragraph beginning at page 29, line 31 to remove the reference to the Bishop reference.